KEN4258: Computational Statistics

Homework Assignment 1 (graded)

Christof Seiler

Posted on March 1 at 20:00 / Due on March 8 at 20:00

Logistics

The final submission in the form of a report should be one single pdf that you can upload to Canvas. One submission per group. The group can have a maximum of 5 members. Keep the report short. The main text should not exceed three pages. You can add an appendix to mention details that you find important. Otherwise, the structure of the report is up to you. You might want to add a link to your GitHub repository with your code. We will select the top three teams. The criteria are clearlity and originality of your work. The three selected groups will each give a 20 minutes presentation followed by a 20 minutes Q&A.

Causal Discovery [100 points]

Your collaborator asks you to make a causal statement about how the variables in the dataset `a1_data.csv` are related to one another.

Follow these steps:

1) Define the assumptions that you would like to make. [25 points]
2) Develop a procedure to estimate the DAG from the data using your assumptions. [25 points]
3) Implement your procedure in R or Python and report your DAG. [25 points]
4) Explain what the limits of your procedure are, both statistically and in terms of possible violations of your assumption with real world data. [25 points]

Load the data and plot histograms.

```
library("readr")
library("ggplot2")
library("tidyr")
a1_data <- read_csv("a1_data.csv")
a1_data
```

```
## # A tibble: 11,672 x 11
##        A     B     C     D     E     F     G     H     I     J     K
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1  3.18  3.03  3.22  2.84  3.39  1.91  2.18  3.46  3.71  3.92  3.57
##  2  3.31  3.13  3.37  2.80  2.53  2.36  2.46  3.39  3.00  3.49  3.75
##  3  3.53  3.55  3.44  2.58  2.73  2.26  2.46  3.45  3.53  3.77  3.25
##  4  3.62  3.83  3.64  2.70  1.73  1.86  2.02  3.57  3.61  3.73  3.33
##  5  3.29  3.21  2.99  2.56  3.01  2.41  2.61  3.33  3.14  3.68  3.87
##  6  3.03  2.48  3.52  2.92  2.65  2.17  2.36  3.63  3.61  3.96  3.73
##  7  3.41  3.47  3.29  3.69  2.83  2.03  2.20  3.77  3.65  3.82  3.22
##  8  3.43  3.09  3.44  3.06  2.86  2.40  2.61  3.52  3.28  3.66  3.27
##  9  3.78  3.70  3.30  2.90  3.24  2.15  2.32  3.50  3.94  4.05  3.37
## 10  3.08  3.24  2.55  3.89  3.26  2.21  3.08  3.18  3.64  3.70  3.97
## # ... with 11,662 more rows
```

```
ggplot(pivot_longer(a1_data, everything()), aes(value)) +
  geom_histogram(bins = 100) +
  facet_wrap(~name, nrow = 3)
```